REPOSITORY FUTURES

Stephen Davison
Caltech Library

Repositories and digital asset management

- Store, manage, preserve/verify, deliver
- Metadata
- Relationships
- Modelling

- Contents
  - Special collections
    - Digitized
    - Born digital
  - Research outputs
    - Published papers, preprints and postprints, technical reports, etc.
    - Data, raw and published

Caltech

Caltech profile

- Major institution, high profile
- High quality research output
- Large publication footprint
- Small institutional size
- Limited Library resources

- Successful institutional repository
- Nascent digital special collections program
  - Digitized collections
  - Born digital archives

Caltech

Principal digital assets

- Archives
  - Archived web sites
  - Born digital assets
    - Email
    - File-based assets (documents, databases, audio, video, …)
    - Other media-dependent assets (DVD, proprietary video formats, …)
    - Oral histories
  - Digitized archival collections
- Research documents
  - Eprints (pre- and postprints, published articles)
  - Research data

Archives
THE CALTECH ARCHIVES

Caltech
coda

Caltech

Principal digital assets

- Archives - 
  - Archived web sites - 
  - Born digital assets
    - Email -  ePADD
    - File-based assets (documents, databases, audio, video, …) - ?
    - Other media-dependent assets (DVD, proprietary video formats, …) - ?
    - Oral histories - 
  - Digitized archival collections - 
- Research documents
  - Eprints (pre- and postprints, published articles, ETDs) - 
  - Research data - 

Caltech

Systems for Digital Asset Management @Caltech

Archival collection management

Digital Library

Email archiving

Institutional Repository

Data Repository

Archived web sites

Software hosting (current)

Local

- ArchiveSpace
- Islandora
- Eprints
- ePADD

Hosted/Service

- Aeon
- Archive-It
- TIND/Invenio ILS

Software hosting (future)

Local

- Islandora
- Eprints
- ePADD

Hosted

- ArchivesSpace/Aeon
- Archive-It
- TIND/Invenio ILS
- TIND/Invenio RDM (Research Data Management)

eprints
repository software

- Open Source
- Shares origins with OAI-OMH
- First and one of the most widely used IR software platforms
- Written in Perl
- Version 3 launched in 2007
- CODA: Caltech Collection of Open Digital Archives
  - 55,000+ research papers
  - Conference papers
  - Masters and PhD theses
  - Oral Histories
  - Campus publications

Caltech

Research Data

- Build on our current repository (EPprints)?
- Acquire another?
- Keep separate, or combine?
- What are the specialized needs of a “data repo”?
  - Greater variety of formats
  - Variety of viewing options, including visualization
- Is the level of curation similar?
- Are the metadata needs similar or different?

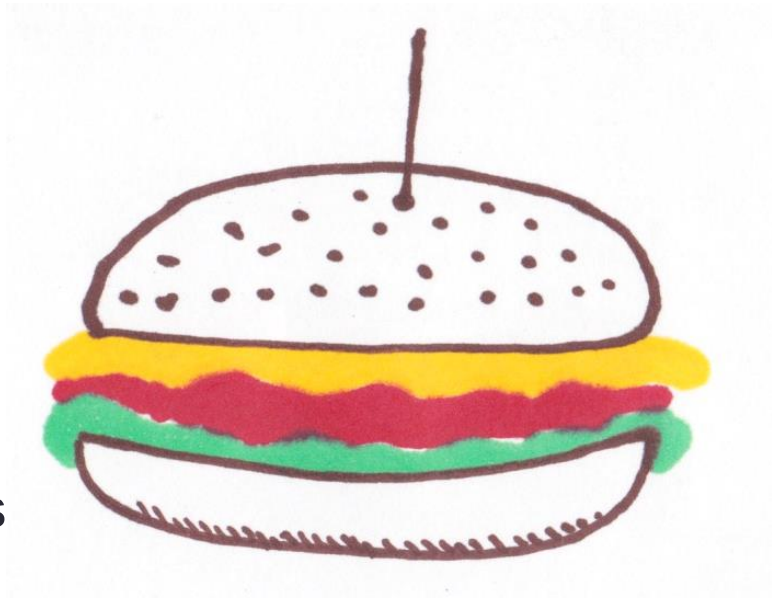
Caltech

Invenio

- Originated at CERN; now widely used
- Open source; Currently co-developed by a consortium including CERN (Geneva), DESY (Hamburg), EPFL (Lausanne), FermiLab (Batavia, IL), SLAC (Menlo Park, CA)
- Highly extensible, modular
- Underlying metadata was MARC, but now JSON-based
- Implementations include:
  - CERN Document Server – cds.cern.ch
    - CERN Open Data Portal – opendata.cern.ch
  - INSPIRE: High Energy Physics Literature DB – inspirehep.net
    - HEPData: HEP Data Repository – hepdata.net
  - Zenodo: open data repository – zenodo.org

Caltech

islandora



- Open source
- Hosted options available
- Components
  - Drupal: CMS, presentation, workflows
  - Islandora: middleware (the plumbing)
  - Fedora: storage and preservation
- Drupal modules / Islandora solution packs
  - Extend functionality in a modular way
  - e.g. “Manuscript Solution Pack” hosts and links:
    - EAD finding aids
    - Digitized manuscripts
    - TEI transcriptions
    - MacCready Collection

Caltech

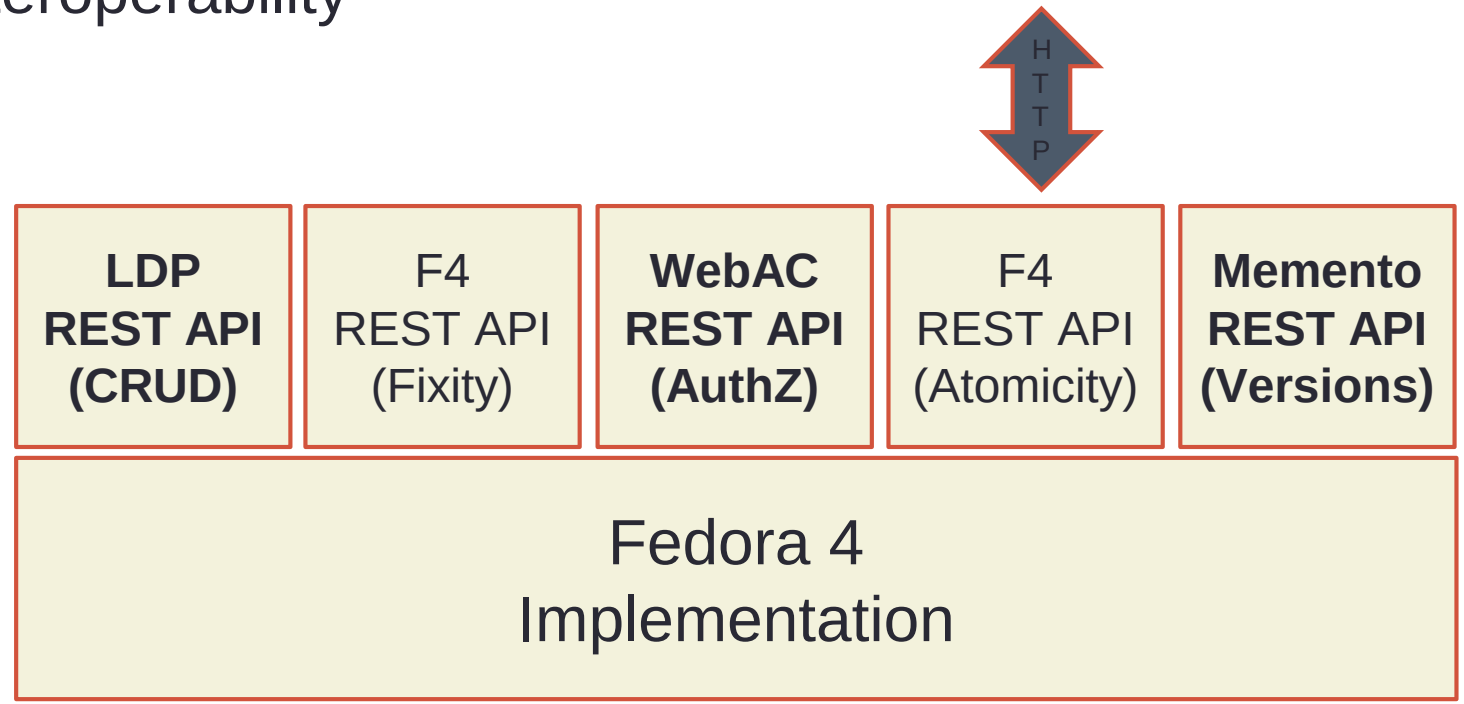
Islandora Fedora 4 implementation

- Islandora CLAW
  - “Claw Linked Asset WebFramework” (self-referential backronym)
- Implements the Portland Common Data Model

Caltech

Fedora 4

- Standards-based RESTful API
- Interoperability



Courtesy Andrew woods, Duraspace

Caltech

REST APIs

- REST: Representational State Transfer
  - Software architectural style of the WWW
  - Not a protocol, but architectural constraints that promote performance, scalability, simplicity of interfaces, modifiability of components, visibility of communication, portability, and reliability
  - Typically communicate over HTTP
  - Interact with external systems as web resources identified as URIs
- API: Application Programming Interface
  - set of routines, protocols, and tools for building software
- RESTful API
  - Uses HTTP requests to GET, PUT, POST and DELETE data

Caltech

Fedora 4 standards

- Linked Data Platform (LDP)
  - CRUD: Create, Read, Update, Delete
  - Very new: W3C Recommendation, 26 February 2015
  - Defines a standard set of techniques to work with RDF resources over HTTP
  - Aligns well with current web technologies, such as REST, Ajax and JSON-LD
- WebAccessControl (WebAC)
  - Implementation of the W3C's still-evolving draft of an RDF-based decentralized authorization policy mechanism
- Memento
  - Standardized architecture for accessing different versions of a doc.
  - Under evaluation by W3C

Caltech

Invenio 3 and Fedora 4

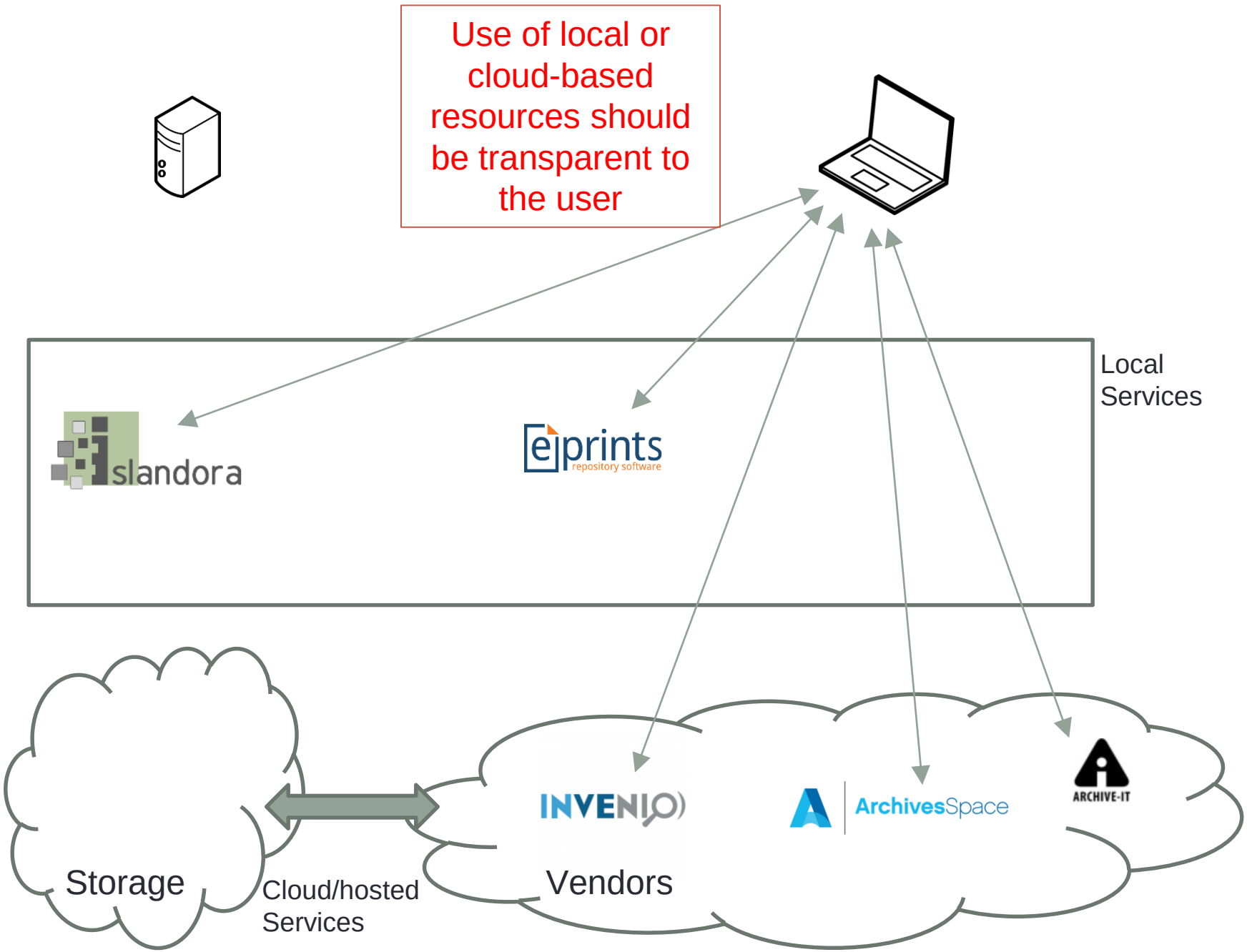
- Highly modular in design
- RESTful API to pass messages between components
- Linked data oriented
  - URIs as identifiers
- Both software stacks have recently been completely rewritten to conform to contemporary standards and to maximize modularity and interoperability

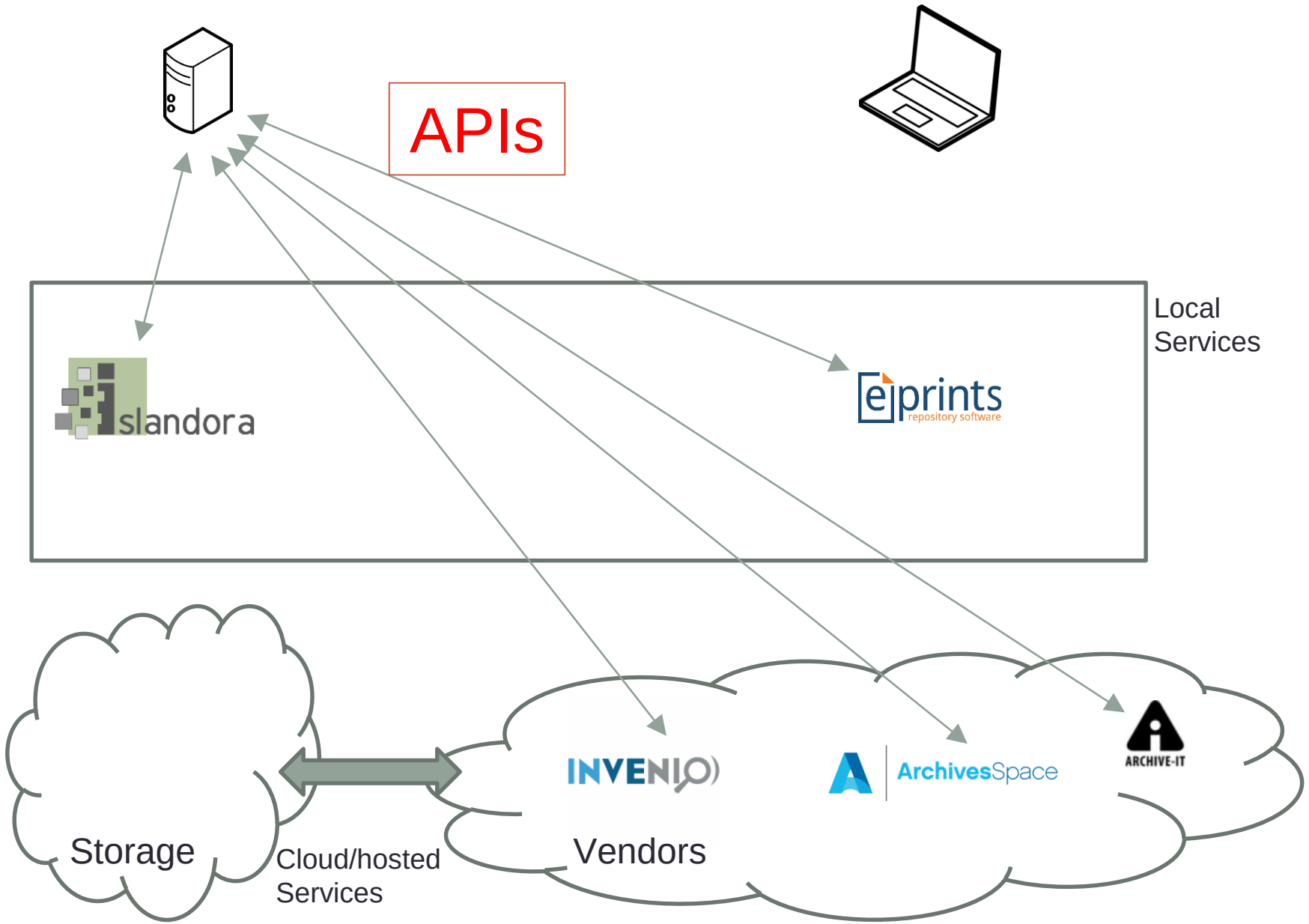
Caltech

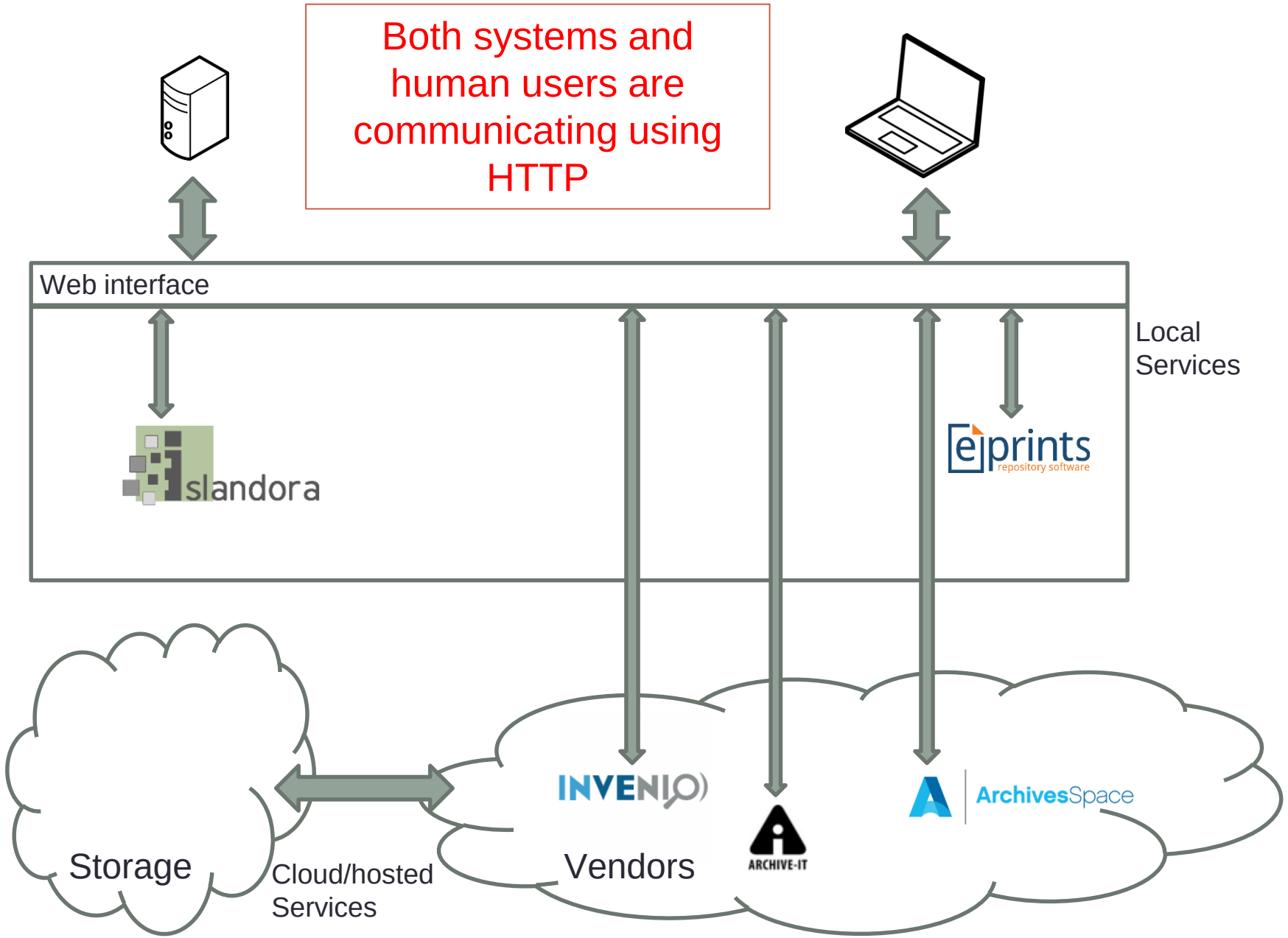
Current Status

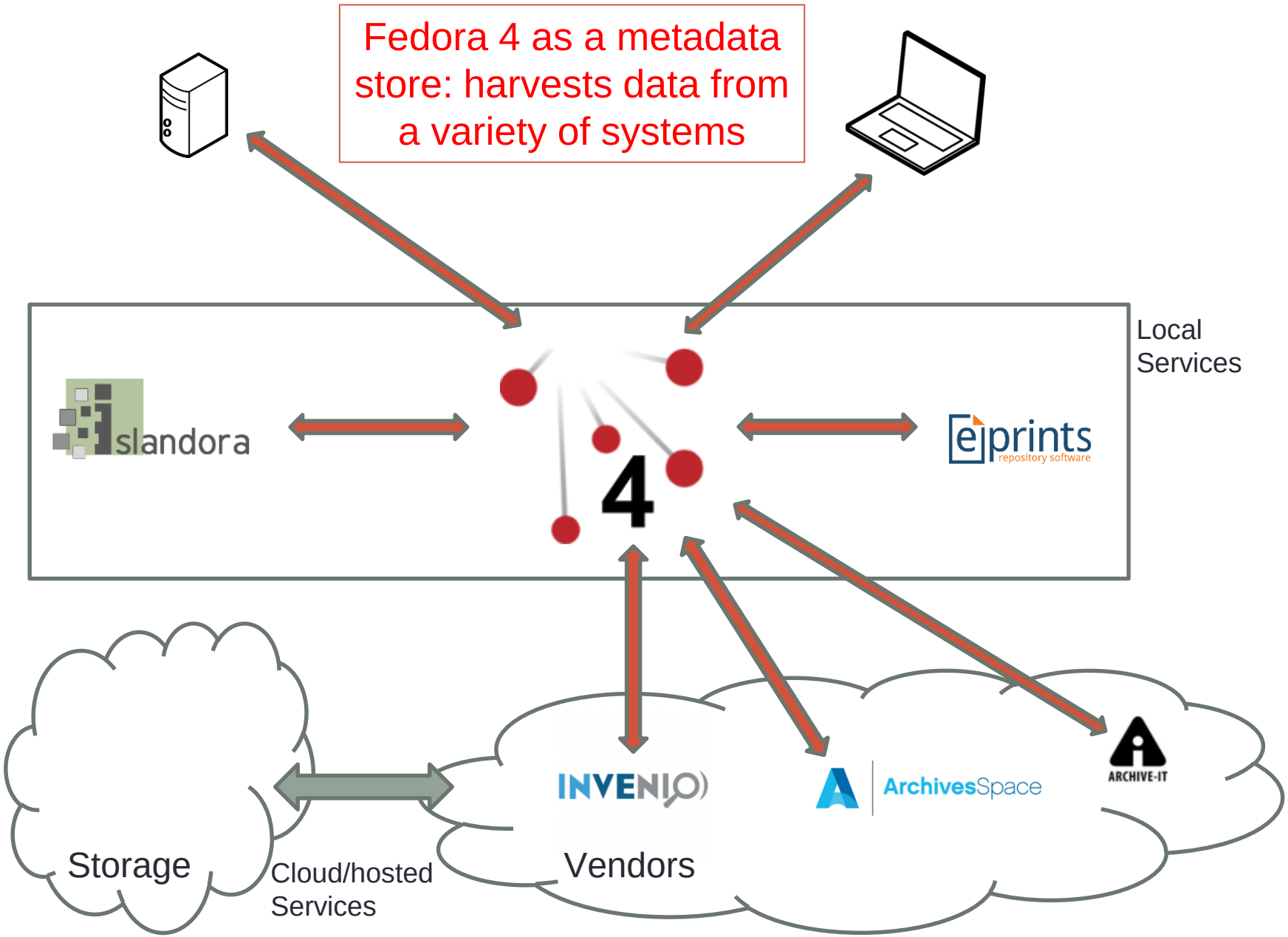
- Multiple repository systems for different content types and purposes
- Inconsistent vocabularies (e.g. names)
- Multiple interfaces

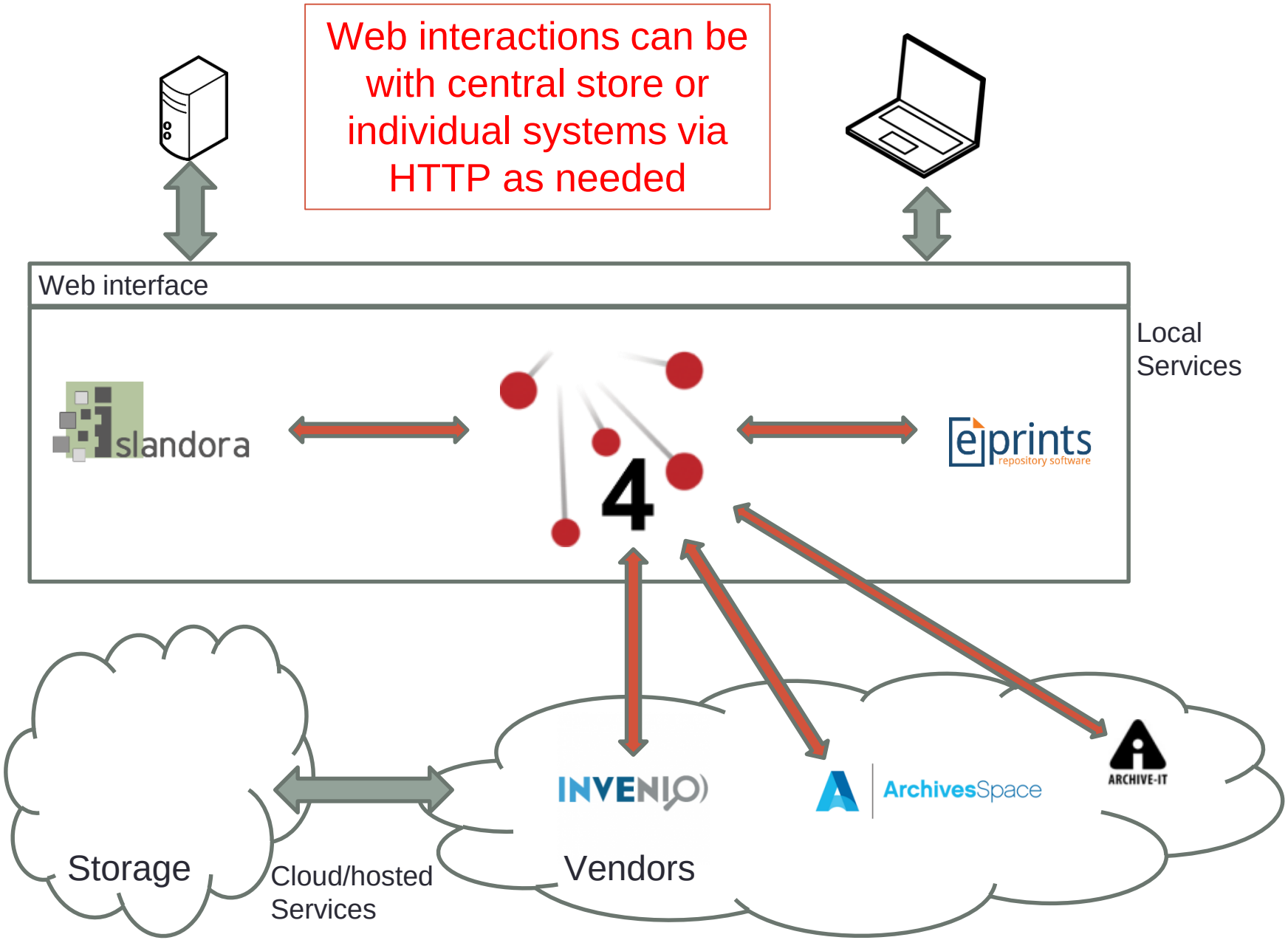
Caltech

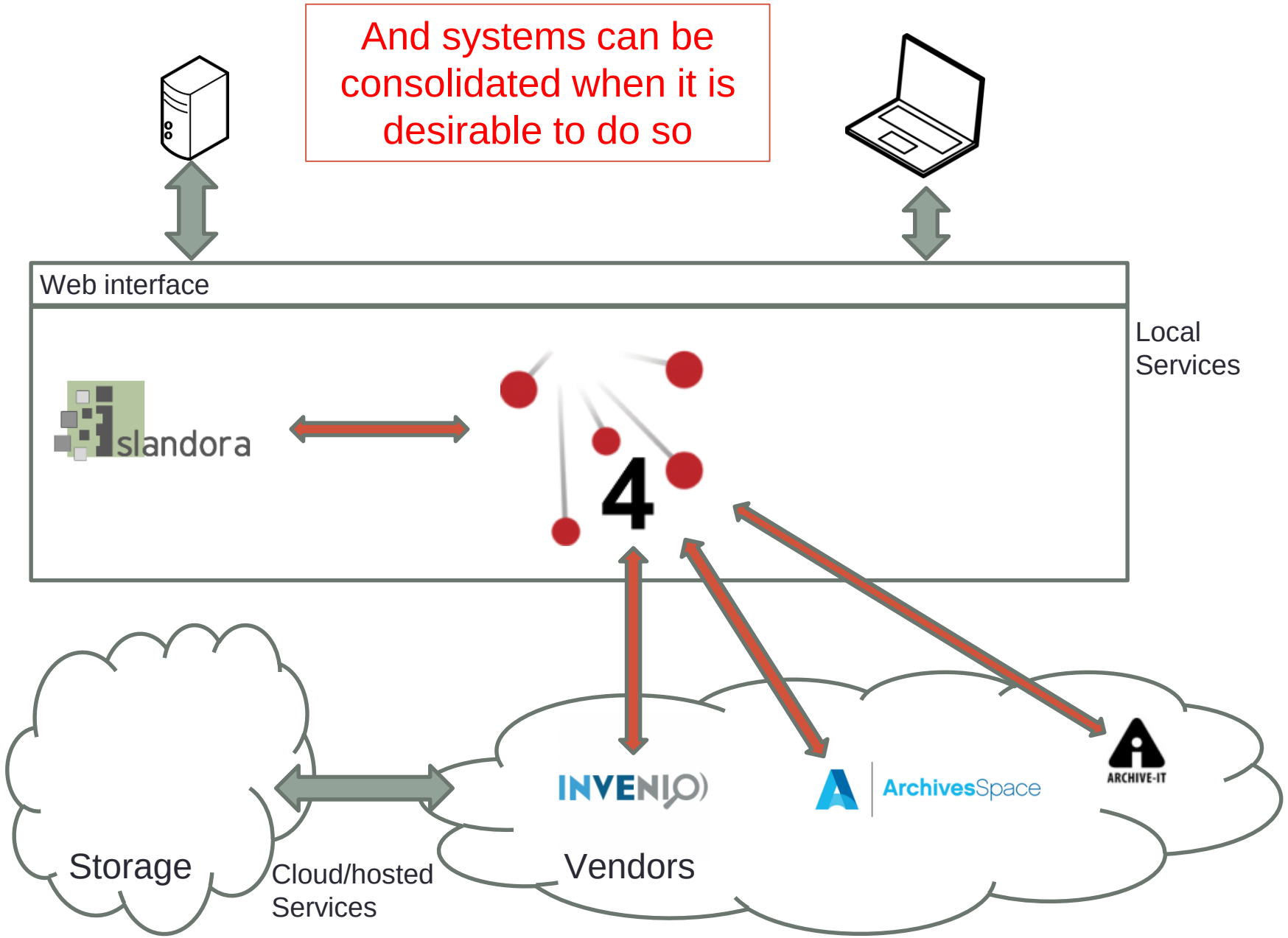

Use of local or cloud-based resources should be transparent to the user
Local Services
Islandora
eprints
repository software
INVENIO
ArchivesSpace
ARCHIVE-IT
Storage
Cloud/hosted Services
Vendors


APIs
Local
Services
islandora
eprints
repository software
Storage
Cloud/hosted
Services
INVENIO
ArchivesSpace
ARCHIVE-IT
Vendors


Both systems and human users are communicating using HTTP
Web interface
Local Services
islandora
eprints
repository software
INVENIO
ARCHIVE-IT
ArchivesSpace
Storage
Cloud/hosted Services
Vendors

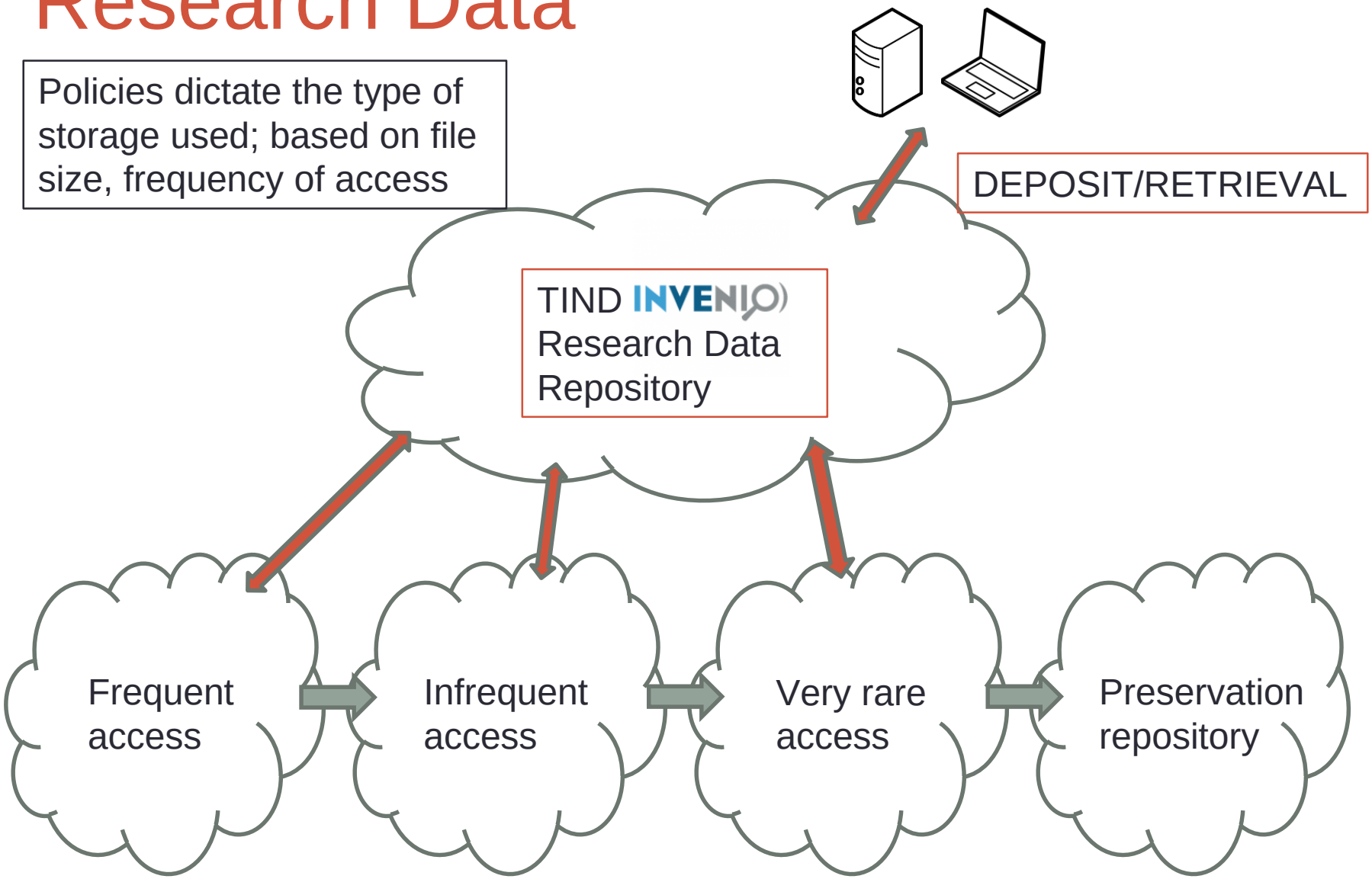

Fedora 4 as a metadata store: harvests data from a variety of systems
Local Services
Islandora
4
eprints
repository software
Storage
Cloud/hosted Services
INVENIO
ArchivesSpace
ARCHIVE-IT
Vendors


Web interactions can be with central store or individual systems via HTTP as needed
Web interface
Local Services
islandora
4
eprints
repository software
Storage
Cloud/hosted Services
INVENIO
ArchivesSpace
ARCHIVE-IT
Vendors


And systems can be consolidated when it is desirable to do so
Web interface
Islandora
4
Local Services
INVENIO
ArchivesSpace
ARCHIVE-IT
Storage
Cloud/hosted Services
Vendors

Research Data

Policies dictate the type of storage used; based on file size, frequency of access



Takeaways

- Outsourcing wherever possible: We are moving away from managing our own software and storage, freeing staff from updating software, backing up data, etc.
- Software and storage that we DO want to manage locally are also migrating to the cloud

Caltech

Takeaways

- Establish and maintain the right number and types of repositories for your institution
- There are downsides to consolidating content and/or metadata; prefer integrated indexing
- Modern programming practices are making distributed systems the norm
- LOD is making the establishment of consistent vocabularies across repositories less urgent
- Be RESTful and achieve good performance, scalability, simplicity of interfaces, modifiability of components, visibility of communication, portability, and reliability

Caltech

Thank you

Stephen Davison
sdavison@caltech.edu